M 1.1:
Running the program gave the desired output of 86.73% accuracy and using time, we measured the elapsed time of the whole python program to be 24.645 seconds.

M 1.2:
The program gave the same output accuracy of 86.73% with an elapsed time of 46.531 seconds.

M 1.3:
The most time consuming kernels are cudaStreamCreateWithFlags, which uses 46.64% of the total time, cudaFree, which uses 28.60% of the total time, and cudaMemGetInfo, which uses 20.87% of the total time.

M 2.1:
With the ece408-high model with a data size of 10000, the model has a correctness of 85.62% with an op time of 11.827049 and a real time of 57.458 seconds. For the ece408-low model with a data size of 10000, the model has a correctness of 62.9% with an op time of 12.185129 and a real time of 55.375 seconds.

Contributions:
Rvarma2: Writing the code for new-forward.h
Stang17: Writing the report

M 3.1:
With the ece408-high model with a data size of 10,000, the model had a correctness of 85.62% with an op time of 0.5585 seconds and a real time of 40.935 seconds. From the nvprof, the forward_kernel took 558.45ms. For the ece408-low model, the model had a correctness of 62.9% with an op time of 0.559 seconds and a real time of 1 minute and 4.698 seconds. The forward_kernel took 559.03ms to complete.

Contributions:
Rvarma2: Writing and debugging code for new-forward.cuh
stang17: Writing code for new-forwawrd.cuh and writing report